APPEN.
VI

APPENDIX VI

Semantic Factoring, Code Construction and Relationships Between Terms

If the index entry "mosquito" is used—together with other terms, of course—to specify the subject matter of a document, its recall is also in order when conducting a search whose scope is defined with the aid of the term "insect". The same will be true of a search defined with the aid of the less generic term "diptera", one of the orders of insects.

The generic terms become available for conducting searching, if the term "mosquito" and related generic terms are represented by the following sets of letters which are recorded as patterns, e.g., of holes in cards. Examples illustrating how this might be done are given below:—

Animals (Kingdom)	ZO
Insects (Class)	ZOKO
Diptera (Order)	ZOKODI
Culicidae (Family)	ZOKODICU
Mosquitoes (Common Usage)	ZOKODICU 34

Animals (Kindom)	ZO
Insects (Class)	ZOKO
Diptera (Order)	ZOKODI
Cecidomyiidae (Family)	ZOKODICE
Phytophaga (Genus)	ZOKODICEFU
Phytophaga destructor (Species)	ZOKODICEFUDE
Hessian fly	

Animals (Kingdom)	ZO
Insects (Class)	ZOKO
Coleoptera (Order)	ZOKOCA
Chrysomelidae (Family)	ZOKOCACO
Diabrotica (Genus)	ZOKOCACODI
Diabrotica undecimpunctata (Species)	ZOKOCACODITU
"spotted cucumber beetle"	
"southern corn rootworm"	

If this type of code is adopted, then encoding a given insect, e.g., the Hessian fly, makes available, for conducting machine searching operations, symbols that indicate that insects genus, family, and order, as well as that is in the class of insects of the animal kingdom. For intelligence purposes, it may not be useful and advantageous

[redacted]
Security Information

to indicate all or even most of this full gamut of generic relationships. Intelligence experts concerned with insects might decide that ability to search for orders or families of insects would not be useful and important. Furthermore, it may turn out that searching operations are not facilitated or expedited by constructing the code so that all insects are indicated as being in the animal kingdom. It might, in fact, prove to make a corresponding simplification in the above codes for insects.

Insects (Class)	KO
Diabrotica (Genus)	KODI
Diabrotica undecimpunetata (Species)	)
"spotted cucumber beetle"	) KODITU
"southern corn rootworm"	)

Insects (Class)	KO
Mosquitoes (Common Usage)	KOCU 34

Insects (Class)	KO
Phytophaga (Genus)	KOFU
Phytophaga destructor (Species)	) KOFUDE
Hessian fly	)

These examples may serve to make the point that the relationship of species to genera to family to order to class to kingdom may (or in part may not) prove useful as a basis for code construction. Other relationships may also prove important. One of the simplest is that of the whole to its parts. An important example is the coding of geographical areas and place names. A simple code for the New England states might be worked out as follows:

United States of America	US
New England	USNE
Maine	USNEME
Vermont	USNEVT
New Hampshire	USNENH
Massachusetts	USNEMA
Rhode Island	USNERI
Connecticut	USNECT

If the above indicates symbols are used, then the act of encoding any one of these six states, makes both the code for New England (USNE) and for United States of America (US) available as a reference point for defining and conducting a search. It should also be noted that it would be easy for the scanning operation to distinguish between USNE standing alone and in combinations such as USNEMA, just as scanning by machine would be able to distinguish, for example, between "cat" as a separate word and the same three letters as found in "catalog", "scatter" or similar words.

[redacted]
Security Information

It is apparent, of course, that such coding of whole-part relationship may not prove sufficient for certain purposes. The codes given above do not provide a means for discriminating between the states as to such characteristic features as industrial development or status as one of the original 13 English colonies. In our simple example, we might attach additional symbols to indicate these features of the New England States. In this case some of the codes would be further extended as indicated below:

United States of America	US
New England	USNE
Maine	USNEME
Vermont	USNEVT
New Hampshire	USNENH,OR
Massachusetts	USNEMA,OR,IN
Rhode Island	USNERI,OR,IN
Connecticut	USNECT,OR,IN

where OR designates one of the thirteen original colonies and IN designates a high degree of industrialization.

In conducting a machine search it will be possible with the new IBM equipment to set up the plug board of the searching machine so that a combination of symbols within a single code are detected. Thus it would be possible to direct the machine to select information relating to all New England states whose code indicates advances industrialization. This would require searching for the combination USNE and IN. If similar codes are worked out for other states in the Union, it would be possible to direct a search to a combination US and IN and select out those items in which one of the highly industrialized states of the union constituted an index entry.

It is important to note in this connection that the desired combination of symbols, e.g., USNE and IN, must be found in the code for a single state. If one element of the combination, say USNE picked up in the code for one state were permitted to interact with the other element namely, IN, in the code for some state outside New England, the door would be opened to the possibility of a false selection. To avoid this the code for each state--or, in general, for each separately coded concept--must begin with a distinctive mark whose detection by the machine will prevent false interactions by resetting those comparator units which have responded to one or more code elements without finding the desired combination within a single code as exemplified by USNENH, OR.

From what has been said, it is perhaps apparent that the purpose of code construction is to render machine searching more effective and efficient. This is accomplished by incorporating in the code general terms such as "industrial" or "original English colony" so they can be used as reference points for defining and conducting a search by automatic equipment. It should be noted in this connection that a large measure of simplification in a coding scheme can be achieved if no important advantage is lost by disregarding certain distinctions which, though perfectly valid from a factual or logical viewpoint, would not be advantageous in imparting needed discriminating power to the machine searching system. If, for example, in dealing with the New England states we are never concerned with an individual state as such but rather with the region as a whole we might decide to employ the code USNE,OR,IN for the region as well as for any regional subdivision, such as one of the states. Furthermore, inclusion of the symbols OR and IN in the code for New England would depend on whether the aspects so indicated are sufficiently important reference points for conducting selecting operations by machine. If, for example, the historical status of some of the New England states as original English colonies is unlikely to be of interest, then the corresponding symbols - namely OR - should be omitted from our codes for the region and its individual states.

This discussion of industrial states illustrates another point in connection with code construction, namely, that certain characteristics are more readily and easily determined than others. No doubt attaches to which of the New England states were among the 13 original colonies. With certain states such as West Virginia an arbitrary decision might be required as to whether this state is to be regarded as highly industrialized. Other things being equal, it is advisable to incorporate in the codes those semantic factors which involve a minimum of arbitrariness in establishing codes.

Enough has been said perhaps to indicate that the effectiveness of a machine indexing system can be greatly increased by devoting care to establishing the most effective possible code for the terminology used for indexing purposes. It is impossible to over-emphasize the importance of simplicity as a desirable element in the final code. In striving to achieve maximum effectiveness in the simplest possible fashion, one of the most important problems in code construction is the selection of general terms to build into the code. In our example such general terms were "United States of America", "New England", "industrial", and "status as one of the 13 original English colonies". This type of general term has come to be spoken of for convenience as a "semantic factor".

During the past year, we have devoted much time to semantically factoring a wide range of scientific and technical terminology to develop a machine indexing system appropriate to the requirements of OSI, in particular, and of the Agency, in general. Before discussing how we went about this, it is well to consider how the semantic factoring technique, when applied to indexing terminology, ties in with the analysis and indexing of documents, on the one hand, and certain machine operations, on the other hand.

To keep the codes for indexing terminology as simple as possible only those semantic factors advantageous as reference points for defining and directing searches should be built into the codes for specific terms. The fact that a given semantic factor may be validly related to a given term does not mean that the factor should be so set up in the code. In fact, care must be exerted to avoid including disadvantageous semantic factors. Consider, for example, the chemical substance, ammonium nitrate, which is used for a variety of purposes. Among these are its use as a fertilizer, as a high explosive, and as a laboratory reagent all of which might be indicated in the code for that substance. From a logical point of view it would be valid to set up the code for ammonium nitrate so as to include "explosive", "fertilizer" and "reagent" as semantic factors. If this were done, a search directed to code symbols for "explosive" would result in the machine selecting all documents for which ammonium nitrate is an index entry even though no mention is made of ammonium nitrate as an explosive. A little reflection will reveal that these additional documents will be troublesome to the extent that the file contains numerous documents dealing with ammonium nitrate but not for explosive purposes. One point illustrated by this example is the fact that the range of subject matter of the documents to be indexed may well have an influence on decisions that are made during code construction. If, for example, a document collection is restricted to papers on explosives, then it may be advantageous to include the semantic factor "explosive" in the code for ammonium nitrate.

Our ammonium nitrate example can also serve to illustrate another important principle in designing codes, and in establishing policy for conducting indexing operations.

Let us imagine that we have before us a document in which ammonium nitrate is mentioned. If ammonium nitrate is mentioned, something will be said about it. Perhaps its physical constants will be listed or the fact that it is being manufactured in a certain plant at a certain place will be stated or its use for one of several purposes will be discussed or described. In any case, if ammonium nitrate is mentioned in a significant fashion, something else must be said about it and consequently there will exist a basis for appropriately qualifying an entry pertaining to ammonium nitrate. This qualification, for example as to its use as a high explosive, is better made at the time of indexing a document than at the time of assigning a code to ammonium nitrate. Such qualification may be accomplished in one of several ways. In the simplest case, the document may refer, for example, to the use of ammonium nitrate as an explosive which latter term is also used as an index entry. Or indexing the document may result in both ammonium nitrate and also some other term, e.g. "torpedo", which has "explosive" as a semantic factor, being used as index entries. The chances that a reasonable amount of indexing will result in the use aspect of a given substance being made available as a reference point for machine searching are--or can be made--so good, that it seems advisable to omit the semantic factor "explosive" from the code for "ammonium nitrate".

Another problem of code construction involves terms whose meanings vary with the context. An example is "low temperature". This term refers to widely different temperature ranges when used to refer to carbonization of coal, to the weather or to research involving liquid helium. The problem of terminology of variable meaning is much less formidable than might appear the case at first glance as association of a term such as "low temperature" with other terminology such as "coal" and "carbonization" tends to accomplish a large measure of definition automatically.

The procedures followed in mass production processing of terms and in their semantic factoring are described in detail in Appendix IX.